DOCUMENT RESUME


| | | |
|---|---|---|
| ED 445 082 | | TM 031 759 |
| AUTHOR | Alexander, Erika D. | |
| TITLE | Using Canonical Correlation To Explore Relationships between Sets of Variables: An Applied Example with Interpretive Suggestions. | |
| PUB DATE | 2000-01-00 | |
| NOTE | 16p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 27-29, 2000). | |
| PUB TYPE | Reports - Descriptive (141) -- Speeches/Meeting Papers (150) | |
| EDRS PRICE | MF01/PC01 Plus Postage. | |
| DESCRIPTORS | Correlation; *Multivariate Analysis | |
| IDENTIFIERS | *Variables (Mathematics) | |



ABSTRACT

Canonical correlation analysis is a parsimonious way of breaking down the association between two sets of variables through the use of linear combinations. As a result of the analysis, many types of coefficients can be generated and interpreted. These coefficients are only considered stable and reliable if the number of subjects per variable is sufficiently large. The first of these coefficients, the canonical correlation, is the bivariate correlation between the composite scores for the two sets of variables. Two additional coefficients, the canonical function and structure coefficients, address the contribution a single variable makes to the explanatory power of the set of variables to which the variable belongs. The communality coefficient explains how useful the variable is in defining the canonical solution. The adequacy coefficient indicates how adequately the analysis represents the total variance in the unweighted set. The extent to which a variable contributes to explaining the composite of the variable set to which the variable of interest does not belong is the index coefficient. A final outcome from canonical correlation analysis is the redundancy coefficient, which indicates the average proportion of variance for variables in one set that is reproducible with the variables in the other set. While the coefficient is easy to calculate, it is not recommended for interpretation in most cases. (Contains 3 tables and 10 references.) (SLD)

ED 445 082

# USING CANONICAL CORRELATION TO EXPLORE RELATIONSHIPS BETWEEN SETS OF VARIABLES: AN APPLIED EXAMPLE WITH INTERPRETIVE SUGGESTIONS

Erika D. Alexander
University of North Texas



TM031759

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, Texas, January 27-29, 2000.

## Introduction

Canonical correlation analysis is a procedure for exploring the relationship between two sets of variables containing two or more variables each. As argued by Baggaley (1981), the multivariate technique is the most general case of the general linear model. It is usually employed because the researcher wants to consider the simultaneous workings of all the variables of interest at once. As noted by Thompson (1984), canonical correlation analysis is appealing because although multivariate in nature, it can be presented in bivariate terms. This is achieved by the calculation of many different coefficients and correlations, each of which answer research questions detailing different aspects of the analysis. This paper presents seven different coefficients that are generated as a result of canonical correlation analysis. Each coefficient is described and interpreted using a practical example.

## Data Example

To illustrate a practical application of canonical correlation, data from a large West coast law school were used to determine the relationship between two sets of variables and their influence on admitted students' decisions whether or not to attend the law school. Eight variables were analyzed using SAS statistical software for the dataset containing 251 observations. The predictor set contained five variables that were measurements of the influence of direct contact of the law school with the admitted students: the admission bulletin, direct mail marketing pieces, official law school forum, law school representative, and visits to the campus. The criterion set contained three

variables that were measurements of outside influences, or influences on the admitted students that are beyond the control of the law school: undergraduate pre-law advisor, attorneys, and parental influences.

According the Barcikowski and Steven's (1975) Monte Carlo study on the stability of the canonical coefficients and correlations, the number of subjects per variable required to achieve reliable results in interpreting the largest canonical correlation should be at least 20/1. When considering the two largest canonical correlations, a ratio ranging from 42/1 to 68/1 should be considered. In this example, the ratio of 31/1 used in this analysis is sufficient to achieve stable and reliable coefficients.

Analysis Results

To obtain an understanding of how well the variables are related to one another, Pearson correlations were calculated among each pair of variables, both within and across sets (see Table 1). The degree to which two predictors correlate is the degree to which they are said to be collinear. The collinearities among the direct contact measurements revealed some moderate values--the largest between FORUM and REPRESENTATIVE, and between ATTORNEY and PARENTS. When considering the between correlations, one notices that ADVISOR is moderately correlated with three of the five direct contact measures and that REPRESENTATIVE is appreciably correlated with each of the outside influence measures.

In canonical correlation analysis, an important measure to consult is the canonical correlation coefficient ($R_c$). According to Thompson (1984), conventional canonical correlation analysis initially begins by "collapsing each person's scores on the variables in

each variable set into a single composite variable." The bivariate correlation between the composite scores for the two sets of variables is the canonical correlation. As explained by Tatsuoka (1971), the total number of possible canonical correlations is equal to $\min(p,q)$ where $p$ is the number of variables in the first set and $q$ is the number of variables in the second set. Therefore, in this example, there are three [$\min(5,3) = 3$] canonical correlations yielded by the analysis.

As seen in Table 2, the two largest canonical correlations, $Rc_1$=0.65 and $Rc_2$=0.27, are both statistically significant at the 0.05 level. One also notices that $R_1$ is larger than any of the between-set correlations. According to Stevens (1996), even though a canonical correlation can be found to be statistically significant, a weak canonical correlation ($R_c$<0.30, $R_c^2$<0.09) may be trivial and of little practical value. Therefore, the researcher may decide a trivial canonical function is not worth interpreting. Because there is such a large decrease in value between $R_1$ and $R_2$ and because $R_2^2$=0.07 is small, only the first canonical function will be interpreted. Results from all three functions are presented in Table 3.

## Result Interpretation

Once a canonical function is identified for interpretation, a number of coefficients may be calculated and consulted to answer various research questions (Thompson, 1984). Of interest to researchers is the contribution a single variable makes to the explanatory power of the set of variables to which the variable belongs. Two coefficients that address this question are the canonical function and structure coefficients. Similar to beta weights in regression, standardized function coefficients are weights applied to the

standardized data, which is summed to create the synthetic variables, or canonical variates (Thompson, 1991). When observing the standardized function coefficients for the direct contact measurements in the first function, one notices that FORUM and REPRESENTATIVE appear to be making the largest contribution, with the other three variables making contributions that are small and similar in size. For the outside influence variables, ADVISOR is making over twice the contribution as either of the other two variables.

As Kerlinger & Pedhazur (1973), Levine (1977), and Meredith (1964) argue, it is important to interpret canonical results based on not only function coefficients, but on structure coefficients as well. Structure coefficients are the bivariate correlations between the predictor variables and the synthetic variable created by the linear combinations, and generally take into account the collinearity, or overlap, of the set of variables. In this example, function and structure coefficients yield similar results. When observing the standardized structure coefficients for the direct contact measurements in the first function, FORUM and REPRESENTATIVE are making the largest contribution, with the other three variables making contributions that are smaller and similar in size. For the outside influence variables, ADVISOR is making the largest contribution. To obtain an estimation of the proportion of variance a variable shares with its canonical composite, the structure coefficient is squared. According to Table 3, FORUM and REPRESENTATIVE account for 80% and 74% of the direct contact variate, respectively, with BULLETIN, MAIL, and CAMPUS accounting for much smaller proportions of the variate. For the outside influence variable set, ADVISOR accounts for 78% of the variate, while ATTORNEY and PARENTS each account for less than 40%.

By summing the squared structure coefficients either across the functions or across the variables within a given function, one obtains the next two coefficients of interest: communality and adequacy. The communality coefficient for a variable (represented by $h^2$) equals the sum of the squared structure coefficients for all the functions and is an indication of what proportion of the variable's variance is reproducible. In other words, how useful the variable was in defining the canonical solution (Thompson, 1984). As seen in Table 3, the communality coefficients indicate that the researcher is not getting as much from the BULLETIN and MAIL variables as from the FORUM, REPRESENTATIVE, and CAMPUS variables.

The adequacy coefficient for a given function is the average of the squared structure coefficients for all the variables in the set and indicates how adequately the analysis represents the total variance in the unweighted set. In this example, the first function has a much larger adequacy coefficient than the other two functions. although the difference is more sizable for the set of variables measuring the direct contact methods.

Also of interest to the researcher is the relationship between the individual variables in one variable set with the canonical variates in the other variable set. In other words, what is the extent to which a variable contributes to explaining the composite, or linear combination of the variable set to which the variable of interest does not belong? The coefficient that addresses this question is referred to as an index coefficient. An index coefficient is the correlation between an unweighted variable in one set and the weighted and aggregated variables in the other set (Thompson, 1984). As seen in Table 3, ADVISOR has the largest index coefficient in the set of direct contact measurements,

and FORUM and REPRESENTATIVE have the largest index coefficients in the set of outside influence measurements.

The final component of canonical correlation analysis is the computation of redundancy coefficients. For a variable set on a function, a redundancy coefficient (*Rd*) is computed by multiplying the adequacy coefficient for the set by $R_c^2$ for the function. It indicates the average proportion of variance for variables in one set that is reproducible with (e.g., redundant with) the variables in the other set. Table 3 shows the *Rd* for each function.

It is often argued that redundancy coefficients should only be interpreted in the "few concurrent validity applications in which both variable sets consist of the same variables" (Thompson, 1991, p.89). Cramer and Nicewander (1979) argued that redundancy coefficients are not truly multivariate "in the strict sense because it is unaffected by the intercorrelaions of the variables being predicted. The redundancy index is only multivariate in the sense that it involves several criterion variables." (p. 43) Therefore, for the heuristic purposes of this paper, *Rd* values were computed and presented; however no interpretations or conclusions will be drawn considering that the research situation from which the data were drawn does not fit the application of redundancy coefficients suggested by Thompson (1991).

Summary

Canonical correlation analysis is a "parsimonious way of breaking down the association between two sets of variables through the use of linear combinations" (Stevens, 1986). As a result of the analysis, many types of coefficients can be generated

and interpretted. These coefficients are only considered stable and reliable if the number of subjects per variable is sufficiently large.

The first of these coefficients, the canonical correlation, is the bivariate correlation between the composite scores for the two sets of variables. Two additional coefficients, the canonical function and structure coefficients, address the contribution a single variable makes to the explanatory power of the set of variables to which the variable belongs. The communality coefficient explains how useful the variable is in defining the canonical solution. The adequacy coefficient indicates how adequately the analysis represents the total variance in the unweighted set. The extent to which a variable contributes to explaining the composite of the variable set to which the variable of interest does not belong is the index coefficient. A final outcome from canonical correlation analysis is the redundancy coefficient, which indicates the average proportion of variance for variables in one set that is reproducible with the variables in the other set. While the coefficient is easy to calculate, it is not recommended for interpretation in most cases.

## REFERENCES


Baggaley, A.R. (1981). Multivariate analysis: An introduction for consumers of behavioral research. *Evaluation Review, 5,* 123-131.

Barcikowski, R., & Stevens, J. P. (1975). A Monte Carlo study of the stability of canonical correlations, canonical weights and canonical variate-variable correlations. *Multivariate Behavioral Research, 10,* 353-364.

Cramer, E.M., & Nicewander, W.A. (1979). Some symmetric, invariant measures of multivariate association. *Psychometrika, 44,* 43-54.

Kerlinger, F.N., & Pedhazur, E.J. (1973). *Multiple regression in behavioral research.* New York: Holt, Rinehart and Winston.

Levine, M.S. (1977). *Canonical analysis and factor comparison.* Newbury Park: Sage.

Meredith, W. (1964). Canonical correlations with fallible data. *Psychometrika, 29,* 55-65.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Tatsuoka, M. M. (1971). *Multivariate analysis: Techniques for educational and psychological research.* New York: Wiley.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation.* Newbury Park: Sage.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development, 24,* 80-95.

## TABLE 1

### Correlations Among the Direct Contact Measures

| | Bulletin | Mail | Forum | Representative | Campus |
|---|---|---|---|---|---|
| Bulletin | 1.0000 | | | | |
| Mail | 0.4630 | 1.0000 | | | |
| Forum | 0.3175 | 0.3902 | 1.0000 | | |
| Representative | 0.2294 | 0.3326 | 0.6370 | 1.0000 | |
| Campus | 0.2615 | 0.1923 | 0.1900 | 0.3128 | 1.0000 |

### Correlations Among the Outside Influence Measures

| | Advisor | Attorney | Parents |
|---|---|---|---|
| Advisor | 1.0000 | | |
| Attorney | 0.2267 | 1.0000 | |
| Parents | 0.2666 | 0.5140 | 1.0000 |

### Correlations Between the Direct Contact and Outside Influence Measures

| | Advisor | Attorney | Parents |
|---|---|---|---|
| Bulletin | 0.2896 | 0.1646 | 0.1831 |
| Mail | 0.3401 | 0.1663 | 0.1791 |
| Forum | 0.5466 | 0.3117 | 0.2858 |
| Representative | 0.4598 | 0.3946 | 0.3502 |
| Campus | 0.1350 | 0.2817 | 0.3018 |

## TABLE 2

### Canonical Correlations

| Function | Rc | $Rc^2$ | Pr>F |
|---|---|---|---|
| 1 | 0.6454 | 0.4165 | 0.0001 |
| 2 | 0.2723 | 0.0741 | 0.0112 |
| 3 | 0.0592 | 0.0035 | 0.8347 |

## TABLE 3

| Variable/ | Function I | | | | Function II | | | | Function III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Coefficient | Function | Structure | Structure2 | Index | Function | Structure | Structure2 | Index | Function | Structure | Structure2 | Index | $h^2$ |
| Bulletin | 0.1407 | 0.4836 | 0.2339 | 0.3121 | -0.0968 | -0.0861 | 0.0074 | -0.0235 | 0.3470 | 0.5409 | 0.2926 | 0.0320 | 0.5339 |
| Mall | 0.1131 | 0.5421 | 0.2939 | 0.3799 | -0.2560 | -0.2260 | 0.0511 | -0.0616 | 0.4638 | 0.4878 | 0.2379 | 0.0289 | 0.5829 |
| Forum | 0.5121 | 0.8944 | 0.8000 | 0.5773 | -0.6108 | -0.2900 | 0.0841 | -0.0790 | 0.1835 | -0.0133 | 0.0002 | -0.0008 | 0.8842 |
| Representative | 0.4263 | 0.8585 | 0.7370 | 0.5541 | 0.4606 | 0.2440 | 0.0595 | 0.0611 | -0.9154 | -0.4080 | 0.1665 | -0.0242 | 0.9630 |
| Campus | 0.1153 | 0.4045 | 0.1636 | 0.2610 | 0.8319 | 0.7854 | 0.6169 | 0.2139 | 0.5008 | 0.4293 | 0.1843 | 0.0254 | 0.9648 |
| Adequacy | | | 0.4457 | | | | 0.1638 | | | | 0.1763 | | |
| Rd | | | 0.1856 | | | | 0.0120 | | | | 0.0006 | | |
| $Rc^2$ | | | 0.4165 | | | | 0.0741 | | | | 0.0035 | | |
| Advisor | 0.7509 | 0.8842 | 0.7818 | 0.5707 | -0.7232 | -0.4586 | 0.2103 | -0.1249 | 0.0496 | 0.0885 | 0.0078 | 0.0052 | 1.0000 |
| Attorney | 0.3481 | 0.6233 | 0.3885 | 0.4023 | 0.4754 | 0.6139 | 0.3769 | 0.1672 | -1.0140 | -0.4844 | 0.2346 | -0.0287 | 1.0000 |
| Parents | 0.2042 | 0.5833 | 0.3402 | 0.3765 | 0.5883 | 0.6399 | 0.4095 | 0.1743 | 1.0083 | 0.5003 | 0.2503 | 0.0296 | 1.0000 |
| Adequacy | | | 0.4264 | | | | 0.2078 | | | | 0.1122 | | |
| Rd | | | 0.2098 | | | | 0.0246 | | | | 0.0006 | | |

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM031759

# REPRODUCTION RELEASE

(Specific Document)

## I. DOCUMENT IDENTIFICATION:

Title: Using Canonical Correlation to Explore Relationships Between Sets of Variables: An Applied Example With Intepretive Suggestions

Author(s): Erika D. Alexander

Corporate Source:

Publication Date:

## II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

| The sample sticker shown below will be affixed to all Level 1 documents | The sample sticker shown below will be affixed to all Level 2A documents | The sample sticker shown below will be affixed to all Level 2B documents |
|---|---|---|
| PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY<br>Sample<br>TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)<br>1 | PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY<br>Sample<br>TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)<br>2A | PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY<br>Sample<br>TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)<br>2B |
| Level 1<br>☑ | Level 2A<br>☐ | Level 2B<br>☐ |
| Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy. | Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only | Check here for Level 2B release, permitting reproduction and dissemination in microfiche only |

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

*I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.*

Sign here, → please

| | | |
|---|---|---|
| Signature: Erika D. Alexander | Printed Name/Position/Title: Erika D. Alexander | |
| Organization/Address: 1529 Mayflower Drive Allen, TX 75002 | Telephone: 972 390 1671 | FAX: |
| | E-Mail Address: edzander@home.com | Date: 01/24/2000 |

## III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

| |
|---|
| Publisher/Distributor: |
| Address: |
| Price: |

## IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

| |
|---|
| Name: |
| Address: |

## V. WHERE TO SEND THIS FORM:

| |
|---|
| Send this form to the following ERIC Clearinghouse: |

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

**ERIC Processing and Reference Facility**
**1100 West Street, 2nd Floor**
**Laurel, Maryland 20707-3598**

**Telephone: 301-497-4080**
**Toll Free: 800-799-3742**
**FAX: 301-953-0263**
**e-mail: ericfac@inet.ed.gov**
**WWW: http://ericfac.piccard.csc.com**